Estimation efficiency in the modeling of dependence structures: an application of alternative copulas to insurance rating

J. D. Woodard[1], N. D. Paulson[2], D. Vedenov[1] & G. J. Power[1]
[1]*Department of Agricultural Economics, Texas A&M University, USA*
[2]*University of Illinois at Urbana-Champaign, USA*

Abstract

This study assesses the performance of several alternative methods for modeling dependence between random variables in the context of pricing an agricultural insurance contract with multiple underlying risk exposures. Simulation methods are used to estimate the sampling distribution of the insurance rates generated under alternative methods. The results indicate significant variability in performance across methods, and contribute to the risk analysis and insurance literatures by quantitatively assessing out-of-sample efficiency and bias trade-off among competing methods for modeling dependence in limited data scenarios.

Keywords: copulas, GRP basis risk, crop insurance rating efficiency, kernel copula, Iman and Conover procedure, Phoon, Quek, and Huang procedure.

1 Introduction

Interest in the modeling of dependence structures and copulas in the fields of risk analysis, financial engineering, and actuarial mathematics has increased substantially in recent years. Copulas [1] offer a more flexible, wider set of tools for modeling dependence structures in probabilistic settings than do more conventional methods—such as Iman and Conover's [2] (IC) or Phoon, Quek, and Huang's [3] (PQH)—but tend to be more computationally intensive, less familiar to those in the field, and more difficult for end-users to implement. Moreover, in many fields little empirical work has actually been conducted to evaluate the performance of alternative copulas and dependence modeling methods in applied settings. The practical importance of properly implementing these methods has increased as a result of the alleged misuse of certain

WIT Transactions on Information and Communication Technologies, Vol 43, ©2010 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/RISK100211

dependence modeling approaches and their role in the global financial crisis of 2008-2009. There is, for example, much concern about correctly modeling systemic risk, which is the likelihood of massive and highly (positively) correlated losses in a system [4].

A particular domain where copulas have not yet been widely adopted but are of great interest is in the rating of insurance for agricultural production exposures [5, 6]. Accurate modeling of the dependence structure is particularly important in agriculture as these insurance covers are typically a function of several related risks. For example, revenue insurance is a function of both output price and quantity of production, and has become the insurance product most widely adopted by farmers in the U.S. Crop Insurance program. While some work in this area exists focusing on the *bias* in rates generated from alternative dependence structures, it is limited and far from comprehensive. Furthermore, to our knowledge, no work has been conducted assessing the out-of-sample *efficiency* of alternative dependence modeling methods in insurance contexts. Efficiency is particularly important in agricultural insurance applications since only a few years of data are generally available to the actuary who estimates rates. Given that insurance ratemaking is primarily a forecasting exercise, assessment of out-of-sample efficiency is of critical importance [7]. Moreover, inefficient rating structures can undermine the integrity of the market through opportunistic behavior driven by asymmetry in information.

We evaluate the bias and efficiency of generated insurance rates under alternative dependence modeling methods and varying data constraints in an inherently out-of-sample framework. The application is to an insurance policy based on crop yields at different levels of aggregation. Specifically, we evaluate a relatively new type of insurance product that indemnifies a producer if the latter's individual yield is less than the yield of the county in which the producer is located. A comprehensive side-by-side comparison of several dependence modeling methods—including the IC and PQH procedures, the Gaussian, Student's-*t*, Frank, Clayton, and Gumbel parametric copulas, a non-parametric kernel copula, and a bootstrap empirical approach—is conducted in a simulation framework which allows for analysis of the estimated rate sampling distribution. To calibrate the simulations, this study employs a large and unique farm-level yield dataset from the Illinois Farm Business and Farm Management database, consisting of 48,568 farm-level yield observations from 1972-2007, as well as matched county-level yield data from the USDA National Agricultural Statistics Service. The results contribute to the risk analysis and insurance literatures by quantitatively investigating the bias-efficiency tradeoff between competing methods for modeling dependence structures. The findings shed light on important questions pertaining to method selection under a variety of data-related constraints typically faced by actuaries and risk managers.

2 The insurance environment

The role of the insurer involves—among other things—pricing risks of potential insureds, as well as underwriting and bearing the risks of actual insureds. The process of pricing insurance, referred to as "insurance rating" or "ratemaking", is addressed by the field of actuarial mathematics. The rating process is crucial to the functioning of the insurance system as a whole. To complicate matters, actuaries are typically faced with data constraints when determining rates, particularly in new markets. Thus, it is important to have an understanding of how different dependence modeling methods perform in small samples in the context of not only expected rate *biases*, but also in terms of out-of-sample rating *efficiency*.

2.1 Insurance policy explanation and motivation

The insurance product employed in this application is a relatively new type of insurance product that insures the excess loss between the farm-level risk exposure and the county-level insurance product. This product is motivated by the nature of the insurance market; in the U.S., a large government-subsidized crop insurance program exists. One product available to producers, known as Group Risk Protection (GRP), is an insurance cover that indemnifies (i.e., pays the insured) based on the county yield. From the producer's perspective, GRP is likely to be effective in protecting against losses that are widespread—such as drought—however, they face a residual "basis" risk in the event that an isolated yield loss event occurs. Thus, the basis risk insurance (*BRI*) product modeled here indemnifies the producer if the individual's loss is greater than the loss payable on the county insurance product.

2.2 Basis risk insurance indemnification structure

The indemnity on the *BRI* is equal to the difference between the county insurance indemnity and the actual producer loss, in cases when the producer loss is greater than the county indemnity. The county indemnity is:

$$I_C = Max\{0, E(Y_C) - Y_C \div Cov_c\} \ , \tag{1}$$

where $E(Y_C)$ is the expected county yield set by the insurer (the insurance guarantee), Y_C is the realized county yield, and Cov_C is the coverage level elected (%) by the producer. Thus, the insurance pays a positive amount whenever the realized yield is less than the guaranteed yield times the coverage level, and has a "disappearing deductible" since the indemnity will equal the total yield guarantee when the realized yield is zero.

The traditional producer-level yield insurance product also pays indemnities when realized yields are less than the yield guarantee, but has a slightly different structure in that it does not have a disappearing deductible. Explicitly, the producer-level indemnity equals

$$I_F = Max\{0, E(Y_F) \times Cov_F - Y_F\}, \tag{2}$$

where $E(Y_F)$ is the expected farm yield, Y_F is the realized farm yield, and Cov_F is elected coverage level. The *BRI* indemnity function is then

$$I_{BRI} = Max\{I_F - I_C, 0\}. \tag{3}$$

In practice, the insurer will attempt to observe I_{BRI} over a large number of cases with similar risk exposures to create an expectation of the cost of offering the insurance, setting the fair insurance rate R equal to expected indemnities. If the joint distribution of producer and county yields is known, then

$$R = \int\int I_{BRI} f(Y_F, Y_C)\partial Y_F \partial Y_C \, . \tag{4}$$

The insurer will typically divide by the size of the underlying exposure (i.e., the "liability"), and then estimate the expected value to derive the rate of expected cost per unit of liability. In this study, the simulation will use exposures with common liability sizes, so we simply refer to it as the rate.

3 Methods for modeling dependence structures

The methods available for modeling dependence structures vary from parametric methods that produce spherical or non-spherical dependence structures, to non-parametric kernel methods, to ad hoc methods of inducing correlation. The agricultural insurance field has been dominated for many years by the use of ad hoc methods such as Iman and Conover's [2] (IC) and more recently some have advocated using the Phoon, Quek, and Huang's procedure [3] (PQH) [8, 9]. Copulas, meanwhile, have received much less attention in agricultural insurance settings.

3.1 The Iman and Conover procedure

The IC procedure [2] is quite widely used in actuarial work and provides a very simple method of generating correlated random uniform variables. Moreover, it is—by actuarial standards and in most applications—very fast and easy to implement. The IC procedure is essentially an ad hoc resorting procedure that uses a random sample of standard normal variates and the Cholesky decomposition of a desired rank correlation matrix to generate correlated standard normal random variates. The correlated standard normal variates are then transformed into correlated random uniform variates via the normal distribution function. The resulting correlated uniforms can then be used to generate correlated random variates with the desired marginal distributions via the inverse distribution method. To implement the procedure, suppose we observe data for M variables for T periods, $X_{T\times M}$ and let $\rho_{M\times M}$ be the estimated rank correlation matrix of $X_{T\times M}$. The IC procedure can be employed to simulate N correlated random uniform variates for M variables with the desired rank correlation $\rho_{M\times M}$ as follows. Letting $\tilde{Z}_{N\times M}$ be a matrix of random (uncorrelated) standard normal variates and $\varsigma(Y_{M\times M}) = C_{M\times M}$ be the Cholesky decomposition of a matrix Y such that $C'C = Y$, we can obtain $N \times M$ correlated uniform random variates as

$$\tilde{U}_{N\times M} = \Phi((\varsigma(\rho)\tilde{Z}')') , \tag{5}$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function (element-by-element).

3.2 Phoon, Quek, and Huang procedure

The PQH procedure [3] is similar to the IC procedure in that it allows for simulation of correlated uniform random numbers with the desired rank correlation. Like the IC procedure, it is both easy to implement and computationally quite fast, but may be more effective at replicating the underlying correlation structure according to Coble *et al.* [8] and Anderson *et al.* [9]. The main insight of the PQH procedure is that a Gaussian process with zero mean and unit variance can be easily simulated with the eigenvalues and eigenfunctions of the covariance function along with a set of uncorrelated standard Gaussian variates by using the Karhunen–Loeve (K-L) expansion representation. Below we provide a simple implementation of the PQH procedure as described by Anderson *et al.* [9] from Phoon *et al.* [3]. Letting $\tilde{\rho} = 2\sin((\pi/6)\rho)$ be the converted Pearson correlation (where again ρ is the estimated rank correlation matrix of sample data X), $\lambda_{M\times 1}$ be a column vector of eigenvalues of $\tilde{\rho}$, $g_{M\times M}$ be a matrix with columns of eigenvectors of $\tilde{\rho}$, $\tilde{Z}_{N\times M}$ be a matrix of random (uncorrelated) standard normal variates, and $\iota_{N\times 1}$ be a column vector of ones, *N* random draws of *M* correlated uniform variables can be generated using the K-L expansion as

$$\tilde{U}_{N\times M} = \Phi(g\sqrt{\lambda}\iota'_N \cdot \tilde{Z}')' , \tag{6}$$

where again $\Phi(\bullet)$ is the standard normal cumulative distribution function (element-by-element).

3.3 Copulas

Copulas allow for the modeling of dependence among marginal variables by directly modeling the joint distribution of standard marginal uniform random variables, where the standard marginal uniforms are the result of the probability integral transforms on the original marginal variables. In actual simulation applications, the copula density is used to generate “correlated” uniform random variates for the variables to be simulated. The resulting “correlated” uniform random variates are then used to generate simulated random variates according to the selected marginal distribution, using the inverse probability transform method. Importantly, copulas do not restrict the underlying uniform marginal distributions to be linearly correlated, but indeed are very general and can model dependence of any type. This is a very important aspect of copulas that separate them from ad hoc methods or more restrictive methods such as the IC or PQH procedures, which typically impose some form of linear or spherical correlation on the underlying uniform marginals. This difference can be very important, particularly in cases where tail dependence or other non-linearities in the covariance structure exist.

3.3.1 Formal definition of Copula and Sklar's theorem

Formally, a copula distribution is a multivariate joint distribution with standard marginal uniform distributions, $C:[0,1]^M \rightarrow [0,1]$, such that $C(\mathbf{u})=0$ if at least one element of $\mathbf{u}\in[0,1]^M$ equals zero, $C(\mathbf{u})=u_m$ if $\mathbf{u}\in[0,1]^M$ has all elements equal to one except element u_m, $C(\mathbf{u})$ is increasing in each element u_m, and its volume $V_C(\mathbf{B})\geq 0$ for every $\mathbf{a},\mathbf{b}\in[0,1]^M$ with $\mathbf{a}\leq\mathbf{b}$ given hypercube $\mathbf{B}=[\mathbf{a},\mathbf{b}]$ whose vertices lie in the domain of C [10]. *Sklar's theorem* states that for any multivariate distribution function, the marginal distributions (modeled via the univariate distribution functions) and the dependence structure (modeled via the copula) can be completely separated. Formally, Sklar's theorem states that for any *M*-dimensional distribution function F with marginal distributions $F_1, F_2, ...F_M$, there exists an *M*-dimensional copula *C* such that

$$F(x_1,x_2,...,x_M)=C(F_1(x_1),F_2(x_2),...,F_M(x_M)). \tag{7}$$

Sklar's theorem has an important corollary that

$$C(\mathbf{u})=C(u_1,u_2,...,u_M)=F(F_1^{-1}(u_1),...,F_M{}^{-1}(u_M)), \tag{8}$$

where $F_m{}^{-1}$ is the inverse distribution function and u_m follows directly by the probability integral transform. Thus, Sklar's theorem shows that copulas can be used to model the underlying dependence structure via the distribution of the probability integral transforms independently of the marginal distributions. Additionally, the use of copulas does not restrict the choice of the marginal distributions in any way. Indeed, any combination of valid marginal distributions can be combined with any given copula, with each unique combination producing a unique multivariate distribution. These features of copulas render them very flexible and powerful tools.

3.3.2 Elliptical parametric copulas: Gaussian and Student's-*t*

The two elliptical parametric copulas assessed in this study are the Gaussian copula and the Student's-*t* copula, which are derived from Sklar's theorem. Accordingly, the Gaussian copula density is defined as

$$C_{\boldsymbol{\rho}}^{Gauss}(u_1,u_2,...,u_M)=\boldsymbol{\Phi}_{\boldsymbol{\rho}}(\Phi^{-1}(u_1),\Phi^{-1}(u_2),...,\Phi^{-1}(u_M)), \tag{9}$$

where $\boldsymbol{\Phi}_{\boldsymbol{\rho}}(\bullet)$ is the standard multivariate normal distribution with Pearson correlation matrix $\boldsymbol{\rho}$ and $\Phi^{-1}(\bullet)$ is the inverse of the standard normal cumulative distribution function. Similarly, the Student's-*t* copula is defined as

$$C_{\boldsymbol{\rho}}^{Student's\,t}(u_1,u_2,...,u_M)=T_{\boldsymbol{\rho},v}(t_v^{-1}(u_1),t_v^{-1}(u_2),...,t_v^{-1}(u_M)), \tag{10}$$

where $T_{\boldsymbol{\rho},v}(\bullet)$ is the standard multivariate Student's-*t* distribution with Pearson correlation matrix $\boldsymbol{\rho}$ and degrees of freedom *v*, $t_v^{-1}(\bullet)$ is the inverse of the standard Student's-*t* cumulative distribution function.

3.3.3 Archimedean parametric copulas: Frank, Clayton, and Gumbel

Archimedean copulas are a relatively flexible class of copulas that can adequately model a wide range of alternative dependence structures, and most have analytical solutions. Following London [10], a simple *m*-dimensional Archimedean copula can take the form

$$C(u_1,u_2,...,u_M)=\varphi^{-1}\left(\sum_{m=1}^{M}\varphi(u_m)\right), \tag{11}$$

where $\varphi:[0,1]\rightarrow[0,\infty)$ is a continuous, strictly decreasing function such that $\varphi(0)=\infty$, $\varphi(1)=0$, and its inverse φ^{-1} is completely monotone on $[0,\infty)$. The Gumbel copula is defined by letting $\varphi(t)=(-\ln t)^{\alpha}$ with $\alpha \geq 0$.

The Clayton copula is defined by letting $\varphi(t)=(t^{-\alpha}-1)/\alpha$ with $\alpha > 0$.

Last, the Frank copula is defined by letting $\varphi(t)=-\ln\dfrac{e^{-\alpha t}-1}{e^{-\alpha}-1}$ with $\alpha \in \Re \setminus \{0\}$.

3.3.4 Non-parametric kernel copula

The kernel copula density is estimated as a multivariate kernel density of uniformly distributed marginals, $U_{T\times M}$, from the probability integral transform of the original data, $X_{T\times M}$, using the empirical distribution. Since the copula has limited support, $[0,1]^M$, the kernel density is truncated using the reflection method. Assuming a product kernel and diagonal bandwidth, the kernel copula density can be written as

$$\hat{c}(u_1,u_2,...,u_M)=$$

$$\frac{1}{Th_1...h_M}\sum_{t=1}^{T}\prod_{m=1}^{M}\left[K_m\left(\frac{u_m-\hat{u}_{t,m}}{h_m}\right)+K_m\left(\frac{u_m+\hat{u}_{t,m}}{h_m}\right)+K_m\left(\frac{u_m-2+\hat{u}_{t,m}}{h_m}\right)\right], \tag{12}$$

where $\hat{u}_{t,m}=F_m(X_{t,m})$ is the empirical distribution of variable *m* [11]. The copula density can then be calculated as the integral of the copula density as

$$\hat{C}(u_1,u_2,...,u_M)=\frac{1}{Th_1...h_M}\sum_{t=1}^{T}\prod_{m=1}^{M}\int_{0}^{u_m}\left[K_m^0+K_m^-+K_m^+\right]dv_m\ . \tag{13}$$

This form has several desirable properties, namely that the marginal copula of a subset of *M* variables can be estimated using only the observations along the dimensions of interest. The result is that adding additional variables to the set involves only a linear increase in the number of computations required to invert each condition marginal copula when sampling from the density.

This study uses a Gaussian product kernel and uses the new bandwidth estimator of Botev *et al.* [12] (BGK). The BGK bandwidth estimator is a non-parametric plug-in approach that involves neither numerical integration nor the normal reference rules that tend to adversely affect other plug-in methods. The BGK bandwidth method was found to perform quite well in terms of both speed and ability to accurately regenerate the underlying data compared to several other bandwidth estimators.

3.3.5 Calibration and simulation of copulas

The parametric copulas are calibrated using the Canonical Maximum Likelihood method (CML). The CML method uses the empirical distribution for each of the *M* variables to convert observed data $X_{T\times M}$ into uniform variates, $\hat{u}_{m,t}$. The copula parameter vector, $\boldsymbol{\alpha}$, is then estimated as

$$\hat{\boldsymbol{\alpha}}_{CML} = \arg\max_{\boldsymbol{\alpha}} \sum_{t=1}^{T} \ln c(\hat{u}_{1,t}, \hat{u}_{2,t}, \ldots \hat{u}_{M,t}; \boldsymbol{\alpha}) . \tag{14}$$

For the elliptical copulas, random draws from the copula can be generated by simply simulating random variates from the standard multivariate distributions with the estimated correlation matrix (and degrees of freedom for the Student's-*t*), and then transforming to uniform marginals using the standard distribution function. That is, letting $\tilde{Q}_{N\times M}$ be a matrix of random standard variates (Gaussian or Student's-*t*) with calibrated parameters $\hat{\boldsymbol{\alpha}}_{CML}$, *N* random draws of *M* correlated uniform variables can be computed simply as

$$\tilde{U}_{N\times M}^{ParamCop} = F(\tilde{Q}_{N\times M}) , \tag{15}$$

where F is the standard distribution function (element-by-element).

The Archimedean copulas can be simulated by first generating independent standard uniform variates, and then inverting the conditional copula density with the proper parameters to generate random draws. Kernel copulas can be simulated in the same way with the exception that there are no parameters to estimate. That is, to generate *N* random draws, $\tilde{U}_{N\times M}^{KCop,ArchCop}$ from an *M*-dimensional from kernel copula or Archimedean copulas, first generate *M* independent vectors of *N* random standard uniform variates, $U_{N\times M}$, and set the first *m*-vector of $\tilde{U}_1 = U_1$. Next, for *m*=2,…, *M,* set $\tilde{U}_m = C_m^{-1}(U_m \mid \tilde{U}_1, \ldots, \tilde{U}_{m-1})$ by solving for $\tilde{U}_m$ using the root-finding equation

$$U_m - C_m(\tilde{U}_m \mid \tilde{U}_1, \ldots, \tilde{U}_{m-1}) = 0 \tag{16}$$

for each element of $\tilde{U}_m = [\tilde{U}_{1,m}, \ldots, \tilde{U}_{N,m}]'$, [13, pg. 184]. The simulated draws are thus

$$\tilde{U}_{N\times M}^{KCop,ArchCop} = [\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_M] . \tag{17}$$

The Frank and Clayton copulas have analytical inverses, so solving the root finding equation is not necessary. The Gumbel, on the other hand, can be inverted numerically. The Gaussian product kernel used for the kernel copula in this study has an analytical conditional distribution (as do many other kernels), although evaluating the inverse remains computationally intensive because at each draw the conditional kernel density must be evaluated several times when solving the root finding equation (eqn. 16).

4 Rate simulation procedure

The simulation employs a Monte Carlo bootstrap resampling procedure to estimate the sampling distribution of the normalized rates under each of the alternative dependence modeling methods. For every combination of $\tilde{T} = \{10, 30, 50\}$ and $\tilde{F} = \{25, 50, 75, 100, 200\}$ (i.e., 15 combinations), J=1,000 bootstrapped yield samples, $\tilde{Y}^{(i)(Boot)}_{\tilde{T}\cdot\tilde{F}\times M}$ consisting of $\tilde{T}\cdot\tilde{F}$ matched county and farm yields, are drawn with replacement from the master dataset of county and farm yields, $Y_{T\times F_t\times M}$ where $t = \{1, 2, ..., T\}$ are the years represented in the dataset, $f_t = \{1, 2, ..., F_t\}$ is the number of farms in the dataset for each year t, $m = \{Farm, County\}$ so that M=2, $\tilde{F}$ is the number of farm yields sampled in every year, and $\tilde{T}$ is the number of years to sample at each iteration i. In order to accurately retain the impact of catastrophic weather events in the sampling distribution when constructing $\tilde{Y}^{(i)(Boot)}$, first a random sample $\tilde{S}_T$ of size $\tilde{T}$ years are drawn with replacement from the T available years of data in Y. Next, for each sampled year $s \in \tilde{S}_T$, a random sample $\tilde{S}_{F,s}$ of size $\tilde{F}$ matched farm and county yields are then drawn with replacement from $Y_{T=s}$. The $\tilde{S}_{F,s}$ samples are then stacked for all $s \in \tilde{S}_T$ to construct $\tilde{Y}^{(i)(Boot)}$.

Marginal uniforms $\tilde{U}^{(i)(Boot)}_{\tilde{T}\cdot\tilde{F}\times M}$ are then estimated via the probability integral transform using the empirical distribution of $\tilde{Y}^{(i)(Boot)}$, and the insurance guarantee is set equal to the sample average of the bootstrapped yield series. Next, each of the dependence modeling methods, $\theta = \{IC, PQH, Kernel, Gauss, t, Frank, Clayton, Gumbel\}$ is fitted to the sample $\tilde{Y}^{(i)(Boot)}$, and a random sample of size N=5,000 Monte Carlo draws of matched county and farm correlated pseudo-random uniform variates, $\tilde{U}^{(\theta)(i)}_{N\times M}$, are then simulated. In order to isolate the comparison and impacts across dependence modeling methods, we use the empirical distributions for the marginals when simulating yields. Thus, the marginal empirical distributions for farm and county yields from $\tilde{Y}^{(i)(Boot)}$ are estimated and then inverted to recover correlated random draws of farm and county yields, $\tilde{Y}^{(\theta)(i)}_{N\times M}$, using the inverse transformation method (this is repeated for each dependence modeling method). $\tilde{Y}^{(\theta)(i)}$ is used to calculate simulated indemnities, $\tilde{I}^{(\theta)(i)}_{(Basis)N\times 1}$, and rates $\tilde{R}^{(\theta)(i)}$ as defined in eq. (1), for every coverage level combinations in $Cov_{Farm} = \{65\%, 70\%, 75\%, 80\%, \quad 85\%, 90\%\}$ and $Cov_{County} = \{65\%, 70\%, 75\%, 80\%, 85\%, 90\%\}$, and for every combinations of $\tilde{T}$ and $\tilde{F}$. Finally, $\tilde{R}^{(\theta)(i)}\ \forall i \in I$ are aggregated to estimate the sampling distribution of $\tilde{R}^{(\theta)}$. The mean and the standard deviation of $\tilde{R}^{(\theta)}$ are reported for each combination defined above. Rates are also calculated from the bootstrapped sample at all iterations. The empirical $E[R]$ is calculated from the original dataset of 48,586 matched yield observations, and is employed as the baseline rate.

Simulations are conducted in MATLAB using the Parallel Computing Toolbox. The IC and PQH procedures as well as the Kernel Copula Monte Carlo simulators were programmed in MATLAB by the authors. The parametric copulas are implemented using the MATLAB Statistics Toolbox.

4.1 Data used for simulation calibration

A large farm-level yield dataset from the Illinois Farm Business and Farm Management database consisting of 48,568 farm yield observations from 1972-2007—as well as matched county-level yield data from the USDA National Agricultural Statistics Service—are used to calibrate the simulations. All farm data were collected from a group of 27 counties in central Illinois with similar risk, soil, and climactic characteristics. Farms were only selected if they contained at least 15 years of data. A feature of agricultural crop yields in this region is that they tend to increase through time due to technological change, improvements in seed biotechnology, and better management practices. Thus, before working with yields it is common to detrend the data [15].

A robust Iterative Reweighted Least Squares Huber *M*-Estimator is employed to estimate trend [14]. The use of robust estimators has gained some popularity in these applications [15], and thus we adopt them here. Using the trend estimate, the trend yield is obtained $Y_i^{Tr} = \hat{\beta}_1 + \hat{\beta}_2 t_i$, where $\hat{\beta}_2$ is the trend yield in year t_i. The detrended yield is thus estimated as $Y_i^{\det} = Y_i + (T - t_i)\hat{\beta}_2$. To reduce sampling variability of the farm-level trend estimates, the county level trend is applied when detrending the farm-level yields. Lastly, to account for differences in the expected yields of each farm and county the data are mean-adjusted such that the resulting detrended means are equivalent for all farms and counties, respectively.

5 Results

Table 1 presents estimated mean rate results by coverage level, and provides a snapshot or rating bias across methods. Compared to the baseline rate, the bootstrap (i.e., empirical copula), Kernel copula, Clayton, and *t*-copula all performed well in terms of bias, although the Clayton tended to be biased downward, while the Kernel and *t* tended to be biased upward. The IC and PQH

Table 1: Mean rate $E(\tilde{R})$ by coverage level combination.

Cov$_C$ (%)	90	90	90	90	90	90
Cov$_F$ (%)	90	85	80	75	70	65
Baseline	1.755	0.765	0.318	0.131	0.053	0.024
Bootstrap	1.755	0.789	0.344	0.151	0.066	0.030
Kernel	2.013	1.018	0.509	0.252	0.122	0.057
IC	2.760	1.585	0.880	0.467	0.235	0.110
PQH	2.713	1.546	0.851	0.447	0.222	0.104
Gauss	2.178	1.108	0.531	0.239	0.101	0.040
Student's *t*	2.075	1.016	0.468	0.205	0.086	0.035
Frank	2.538	1.464	0.835	0.464	0.249	0.127
Clayton	1.567	0.573	0.179	0.053	0.016	0.005
Gumbel	2.651	1.511	0.838	0.447	0.229	0.111
Cov$_C$ (%)	75	75	75	75	75	75
Cov$_F$ (%)	90	85	80	75	70	65
Baseline	3.679	2.112	1.121	0.542	0.236	0.094
Bootstrap	3.522	2.030	1.096	0.548	0.254	0.111
Kernel	3.600	2.141	1.226	0.670	0.348	0.172
IC	3.802	2.338	1.390	0.787	0.418	0.207
PQH	3.789	2.324	1.376	0.775	0.409	0.200
Gauss	3.625	2.155	1.217	0.638	0.305	0.133
Student's *t*	3.602	2.120	1.174	0.595	0.272	0.112
Frank	3.792	2.347	1.415	0.821	0.453	0.236
Clayton	3.505	1.983	1.005	0.425	0.144	0.042
Gumbel	3.789	2.325	1.380	0.779	0.414	0.206

Note: Table presents mean rates from the rate simulations for each different coverage level combination and method. Due to space considerations, some coverage level combinations are excluded.

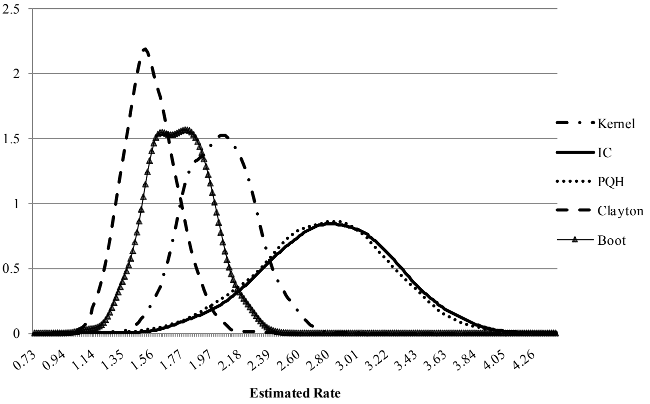

Figure 1: Rate distributions ($\tilde{F}$=100, $\tilde{T}$=30), Cov$_C$=90%, Cov$_F$=90%.

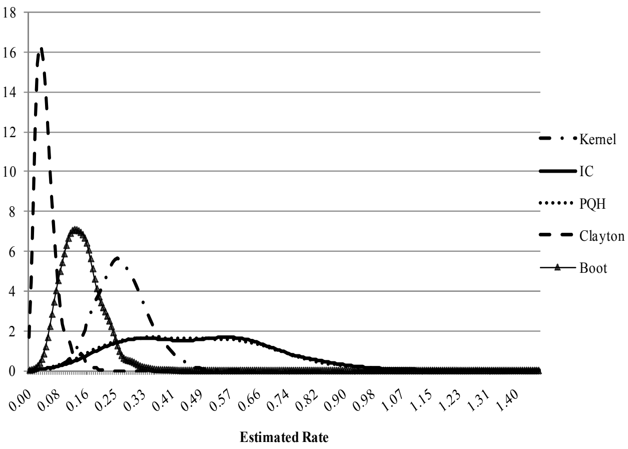

Figure 2: Rate distributions, $\tilde{F}$=100, $\tilde{T}$=30, Cov_C=90%, Cov_F=75%.

tended to perform most poorly, overestimating the rate by over 50% for the 90%/90% coverage product, and by over 350% for the 90%/65% product.

To assess out-of-sample rate efficiency we evaluate the distribution of rates under each method. Figures 1 and 2 present rate distributions for five selected methods, and Table 2 presents rate standard deviations for several coverage levels. The rate distributions in Figures 1 and 2 can be interpreted as the distribution of the rates an insurer would estimate under adoption of a given dependence modeling method given the bootstrapped/sampled data they observe; or, as the distribution of indemnities paid by the insurer over a given horizon for the specified portfolio of insurance. At low coverage level combinations, the rates tend to be highly volatile across methods, and IC and PQH tend to perform much worse than the other candidate methods. For example, Figure 1 illustrates that not only do the PQH and IC perform nearly identical, but are also the most biased and inefficient (90%/90% coverage), and Figure 2 shows that their performance relative to the other methods deteriorates substantially as the coverage level decreases (90%/75%). The Kernel, Bootstrap, *t,* and Clayton all tended to perform quite well in terms of efficiency, although at low coverage levels the Clayton tended to generate unacceptably low rates (Figure 2). Interestingly, the Bootstrap/empirical copula tended to perform best overall in terms of both bias and efficiency, regardless of coverage level.

The sampling procedure also allows for assessment of increases in rate efficiency as more farms or more years of data are observed by the insurer, with the results suggesting that having more *years* of data is preferred relative to having more *farms in each year*. The reason for this finding is that farm yields can be highly correlated in any given year due to spatial correlation in weather, resulting in only minimal additional information from the addition of more observations in a given year. For example, referring to the Kernel copula rate

Table 2: Rate efficiency $\sigma(\tilde{R})$, Cov$_C$=90%, Cov$_F$=90%.

$\tilde{T}$	*10*	*10*	*10*	*10*	*10*	*30*	*30*	*30*	*30*	*30*	*50*	*50*	*50*	*50*	*50*
$\tilde{F}$	*25*	*50*	*75*	*100*	*200*	*25*	*50*	*75*	*100*	*200*	*25*	*50*	*75*	*100*	*200*
$\tilde{F}\times\tilde{T}$	*250*	*500*	*750*	*1000*	*2000*	*750*	*1500*	*2250*	*3000*	*6000*	*1250*	*2500*	*3750*	*5000*	*10000*
Baseline	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bootstrap	0.462	0.394	0.397	0.370	0.359	0.261	0.232	0.213	0.215	0.200	0.201	0.180	0.179	0.160	0.161
Kernel	0.532	0.423	0.415	0.403	0.365	0.305	0.262	0.237	0.239	0.237	0.254	0.217	0.217	0.196	0.196
IC	0.802	0.740	0.784	0.731	0.738	0.493	0.475	0.494	0.475	0.478	0.405	0.403	0.410	0.395	0.396
PQH	0.782	0.727	0.743	0.725	0.704	0.502	0.468	0.491	0.462	0.460	0.385	0.380	0.385	0.386	0.384
Gauss	0.556	0.488	0.509	0.478	0.452	0.331	0.311	0.321	0.302	0.307	0.273	0.254	0.268	0.244	0.243
Student's *t*	0.519	0.466	0.454	0.442	0.421	0.306	0.279	0.289	0.270	0.269	0.251	0.247	0.233	0.225	0.219
Frank	0.624	0.565	0.606	0.553	0.540	0.405	0.381	0.399	0.379	0.384	0.337	0.330	0.344	0.325	0.318
Clayton	0.420	0.359	0.324	0.338	0.315	0.224	0.209	0.191	0.187	0.184	0.186	0.173	0.166	0.145	0.155
Gumbel	0.723	0.686	0.717	0.664	0.659	0.467	0.446	0.464	0.438	0.444	0.371	0.370	0.370	0.361	0.375

Note: Table presents estimated rate standard deviation calculated over all *I=1,000* simulated rates for each bootstrap sample size, for the 90%/90% coverage level product.

efficiency results for the in Table 2, the rate standard deviation when sampling 10 years and 200 farms (0.365) is substantially higher than when sampling 30 years and 25 farms (0.305), even though in the latter case fewer actual yields are sampled overall (2000 vs. 750). Furthermore, the marginal value of adding more farms in a given year appears to diminish quickly in the farm sampling size. For example, the efficiency gains in Table 2 are high as the number of sampled farms goes from 25 to 50 when sampling 10 years, (0.462 to 0.394 for the Bootstrapped rate), but only small efficiency gains are realized when going from 50 to 200 farms (0.394 to 0.359).

6 Conclusion

This study assessed the performance of several alternative methods for modeling dependence between random variables in the context of pricing an agricultural insurance contract with multiple underlying risk exposures. Simulation techniques were used to estimate the sampling distribution of the insurance rates generated under each method in order to assess the bias and efficiency of the rating structures implied by each method. Simulations were also conducted across several different resampling sizes to investigate the effect of data availability on rating efficiency. For this particular application, we find that the bootstrapping method, kernel copula, and Clayton copula all perform quite well given the structure of the data, but that the Clayton copula often generated rates that were unacceptably low. The Frank and Gumbel copulas tended to underestimate the tail dependence and produce somewhat biased and inefficient ratings for this product. The PQH and IC procedures performed similarly and very poorly. Indeed, they produced the most biased and inefficient results.

These results have several implications for insurance rating and risk analysis, and contribute to the literature by quantitatively assessing the bias and out-of-sample efficiency among several competing methods for modeling dependence structures. A better understanding of these bias and efficiency characteristics in empirical settings is useful for researchers, actuaries, and risk managers working on a range of insurance problems, particularly when data are limited. They highlight the importance of the dependence modeling technique chosen by the practitioner, point out the restrictive nature of parametric copulas, and illustrate some of the benefits of the kernel copula. This study also calls into question the results of recent studies by Anderson *et al.* [9] and Coble *et al.* [8] which purport to find significant differences in the performance of the PQH and IC procedures in the context of rating revenue and whole farm-insurance products, and which recommend that the RMA-USDA convert their systems to use the PQH instead of IC. We find that there is no significant difference between PQH and IC in this application, and also find no reason to expect that this would not be the case of rating revenue products. Nevertheless, additional work is needed to assess revenue product rating in comprehensive manner similar to that employed here. Furthermore, our results suggest that even compared to copulas with similar shape characteristics, the PQH and IC procedures perform much worse in terms of out-of-sample efficiency. This has important implications for the U.S. Federal

Crop Insurance program given its reliance on the IC method for rating revenue insurance.

Further research is needed to assess the performance of these and other methods in related contexts. For example, there also exist revenue insurance product analogs to the yield-only basis risk insurance product investigated in this study. Future work could also extend the simulation techniques employed to estimate the evolution of the insurance market and underwriting performance with uncertain rates under alternative methods.

References

[1] Nelsen, R.B., *An Introduction to Copulas*, Springer: Berlin and NY, 2006.

[2] Iman, R.L., Conover, W.J., A distribution-free approach to inducing rank correlation among input variables. *Communication in Statistics-Simulation and Computation,* **11(3)**, pp. 311-334, January, 1982.

[3] Phoon, K., Quek, S.T., and Huang, H., Simulation of non-Gaussian processes using fractile correlation. *Probabilistic Engineering Mechanics*, **19**, pp. 287-292, 2004.

[4] Kaufman, G.G. and Scott, K.E., What is systemic risk, and do bank regulators retard or contribute to it? *The Independent Review*, **7(3)**, pp. 371-91, 2003.

[5] Vedenov, D., Application of copulas to estimation of joint crop yield distributions, *American Agricultural Economics Association Annual Meeting*, Orlando, FL, July 27-29, 2008.

[6] Zhu, Y., Ghosh, S.K., and Goodwin B.K., Modeling dependence in the design of whole farm insurance contract: A copula-based model approach. *American Agricultural Economics Association Annual* Meeting, Orlando, FL, July 27-29, 2008.

[7] Norwood, B., Roberts, M., and Lusk, J., Ranking crop yield models using out-of-sample likelihood functions, *American Journal of Agricultural Economics*, **86(4)**: pp. 1021-1043, 2004.

[8] Coble K.H., Harri A., Anderson, J.D., Ker, A.P., Goodwin, B.J., *USDA Risk Management Agency Review of County Yield Trending Procedures and Related Topics*, February 18, 2008.

[9] Anderson, J.D., Harri, A., and Coble, K.H., Techniques for multivariate simulation from mixed marginal distributions with application to whole-farm revenue simulation. *Journal of Agricultural and Resource Economics,* **34(1)**, pp. 53-67, 2009.

[10] London, J., *Modeling Derivatives Applications in Matlab, C++, and* Excel, Monte Carlo and New York, p. 69, 2006.

[11] Charpentier, A., Fermanian, J.D., and Scaillet, O., *The Estimation of Copulas: Theory and Practice*, Rank, J. (ed). Risk Books: London, 2007.

[12] Botev, Z.I., Grotowski, J.F., and Kroese, D.P., Kernel density estimation via diffusion. *Annals of Statistics*. Accepted for publication, 2010.

[13] Cherubini, U., Luciano, E., and Vecchiato, W., *Copula Methods in Finance*, Wiley Finance Series, John Wiley & Sons: Hoboken, NJ, 2004.

[14] Fabozzi, F.J., Kolm, P.N., Pachamanova, D., *Robust Portfolio Optimization and Management*, The Frank J. Fabozzi Series, John Wiley and Sons, Inc.: Hoboken, New Jersey, June 2007.

[15] Ramirez, O.A., Misra, S.K., and Nelson, J., Efficient estimation of agricultural time series models with nonnormal dependent variables. *American Journal of Agricultural Economics*, **85(12)**, pp.1029-1040, November 2003.